REPORT RESUMES

ED 003 898 48

APPLICATION OF LEARNING THEORY TO PROBLEMS OF SECOND LANGUAGE ACQUISITION, WITH PARTICULAR REFERENCE TO RUSSIAN.

BY- SUPPES, PATRICK
STANFORD UNIV., CALIF., INST.FOR MATH.STUD.SOC.SCI

REPORT NUMBER NDEA-VI-93 PUB DATE 31 AUG 65
CONTRACT OEC-5-14-013
EDRS PRICE MF-$0.18 HC-$3.76 94P.

DESCRIPTORS- *LEARNING PROCESSES, *RUSSIAN, ENGLISH, STANFORD, CALIFORNIA

SUPPES' 1963 MODEL WITH TWO PROCESSES, LEARNING AND FORGETTING, WAS USED IN AN ATTEMPT TO PREDICT THE OPTIMAL "LESSON SIZE" FOR TEACHING A LIST OF INDEPENDENT ITEMS. THE EXPERIMENT REPORTED IN THIS DOCUMENT WAS A TEST OF THIS PREDICTION IN A RUSSIAN-ENGLISH VOCABULARY LEARNING SETTING. (LP)

ED003898

Somehow related to (93)

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

FINAL REPORT

APPLICATION OF LEARNING THEORY TO PROBLEMS OF SECOND LANGUAGE ACQUISITION, WITH PARTICULAR REFERENCE TO RUSSIAN

CONTRACT OE 5-14-013

PARTICK SUPPES

The Research Reported Herein was Performed
Pursuant to a Contract with the United
States Office of Education, Department of
Health, Education, and Welfare
(under provisions of the Language
Development Program, NDEA)

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES,
STANFORD UNIVERSITY,
STANFORD, CALIFORNIA

September 1, 1964 to August 31, 1965

Date of Report: August 31, 1965

Table of Contents

- Introduction.............................. 1
- Vocabulary learning experiment......... 2
    - method................................ 4
    - results and discussion............... 7
    - summary............................... 16
    - footnotes............................. 17b
    - tables 1-8............................ 18-23b
    - item sequencing model................ 24
    - item analysis........................ 25
- Grammar learning experiment............ 30
    - method................................ 34
    - results, classification test....... 47
    - results, production test........... 64
    - summary............................... 69
    - footnotes............................. 73
    - tables 1-10........................... 74-83
- References................................ 84
- Appendix A................................ 86
- Appendix B................................ 89

During the past fiscal year our research on the acquisition of Russian as a second-language focused on two problems, grammar learning and vocabulary learning. Major emphasis was accorded to variables affecting the acquisition of Russian syntax; the outcome of this research is reported in the second half of this paper. Prior to that study we completed the empirical test of predicted relationships between vocabulary block-size and learning rate; the findings appear in the first half of this report, together with an analysis of linguistic correlates of vocabulary item difficulty. The syntax and vocabulary studies both evince a concern with how item presentation order influences learning. A theoretical analysis of some aspects of this problem is inserted between the two major sections of the report.

As indicated by the Table of Contents, the report of each main study begins with an introduction, which sets forth the specific problem and objectives. Next occur the description of method, results and discussion. Finally, each ends with a summary or abstract, and these are submitted in lieu of a single overall summary. Plans are underway for publication of a research monograph on the findings from FY 1961-62 to date. We anticipate that the manuscript will be completed by the end of the calendar year. Some of the material in the present report is excerpted from those pre-publication drafts; this explains occasional references such as "in the preceding chapter".

Optimal Block Size in a Russian Vocabulary Learning Experiment

Suppes (1963) model with two processes, learning and forgetting, proposes to predict the optimal "lesson size" for teaching a list of independent items. The experiment reported here is a test of this prediction in a Russian-English vocabulary learning setting. Before describing the experiment it is necessary to review briefly the theoretical formulation. Suppose that the number of items to be learned equals X, and that X can be factored into integers Y and Z. Suppose also that the experimenter wishes to reinforce each item k times. Then a presentation order with block size k is: (1) items 1 thru Y presented in sequence with one reinforcement each, then k-1 repetitions of this sequence; (2) iterate with items Y+1 through 2Y; (3) continue iterating through successive blocks of Y items, i.e. of size Y; (4) give a test cycle to measure learning under block size Y. Since Y can vary from one list to another (e.g., if X = 72, Y may be any factor, including 1, 2, 18, 36, or 72), we may ask what value of Y is theoretically optimal with respect to performance on the post-test. According to Suppes' two-process model, the optimum is either Y = 1 or Y = X. A related prediction, to be tested by the present experiment, is that test performance should be a monotonic function of Y. That is, with X = 72 we selected three values of Y - 18, 36, and 72. The prediction now is that the only possible rank-orders from highest to lowest test proportion correct are: 18, 36, 72 and the reverse. In other words, size 36 must yield intermediate performance.

Our previous experiments gave tentative support for the prediction, but findings were inconclusive because of sizeable intersubject differences in learning. The present experiment attempted to control this variability by running all subjects under all block sizes, instead of only one. Since the three sizes 18, 36, and 72 can be ordered in 3! = 6 ways, it was necessary to run six groups of subjects in order to counterbalance the order of occurrence of the sizes. Hence altogether each subject received three different sets of 72 items. In order to produce a reasonable degree of learning, each item within a set received four reinforcements (by the anticipation method). The practical schedule was to present the three sets, each followed by its test, on three successive days. A secondary aim of the experiment involved determining the amount of retention. One recall test of all 216 was given on the day after the third acquisition set, with a follow-up test one week later.

Subjects. Subjects were from the same experimental population as in the preceding study. None had participated in that experiment or in any of our earlier vocabulary learning investigations.

Experimental groups. There were six groups of eight subjects each, one group corresponding to each of the 3! = 6 possible presentation orders of block sizes 18, 36, 72. We denote the groups according to their first block size, as shown in Table I.

Table I here

Assignment of subjects to groups was based on the time of day when the student agreed to participate. All members of a group were run concurrently.

Sessions 1-5 for groups 18A, 36A, and 72A spanned one five-day period; the other groups were run during the next five-day interval. For all groups, twenty-four hours elapsed between sessions, except that one week intervened between Sessions 5 and 6.

Materials. A total of 338 words were used, 216 in the experiment proper and 126 in the pre-training session. The 216 comprised 96 from Josselson's previously described Lists 1 and 2, plus 20 from his lower-frequency Lists 3 and 4. There were 87 nouns, 45 verbs, 4 pronouns, 33 adjectives, 40 adverbs, 3 conjunctions and 4 prepositions. Pre-training words were drawn from Lists 5 and 6. Words that closely resembled the 216, either in pronounciatior or meaning, were excluded from the pre-training list.

The 216 words were divided into three sets of 72; division was subject to two constraints. First, the number of nouns was the same in all three sets, and likewise for the verbs and adjectives. In order to fulfill the second requirement the 216 words were chosen from among those previously used in the 100/300 study. The 216 were divided into three sets such that the total errors per set, based on the group 300 data, was equal across sets. Constructing homogeneously difficult sets was intended to permit comparisons across sets for a fixed block-size. One set of 72 was used in Session 2 for all groups, another set in Session 3, and the third set in Session 4.

Procedure. Subjects were run in groups in a soundproof room. Each was supplied with a numbered answer sheet containing blanks in which to write the English equivalent of the auditorily presented Russian word. After hearing the Russian, the subject had 5 seconds in which to respond. Two seconds elapsed between the English reinforcement (i.e. the correct answer played over the tape) and the next Russian stimulus word.

Instructions and items were played over the Ampex recorder described in a previous report. First there were general instructions, identical to those previously used, informing the subject of the nature of the task, the event sequence on a trial, and that a written response must be exactly right in order to be graded as correct. Next there were instructions specific to the block size. These were added so that the subject would know that a final test would occur. The following specific instructions for size 18 are illustrative of those for all sizes:

> You will hear 18 Russian words, one at a time. Then you will hear the same 18 three more times. Next the procedure will be repeated with a second set of 18, and so on, until you have had 72 items, four times each. Finally, you will have a test of all 72. Do you have any questions?

Session 1 - Pretraining. This session was included in order to minimize possible warm-up effects at the outset of the training phase. Pretraining was an abridged form of the subsequent training, except for a modification to insure that pretraining did not favor a particular block size. Each subject had block size 18 with one set of words, size 36 with a different set of words, and size 72 with a third set. But in order to demonstrate all three sizes in the same session, only one block of each size was presented (instead of, say, the normal four in block size 18). Hence a total of 18 + 36 + 72 = 126 words were used during pretraining. After two random cycles through the 18, there followed two cycles through the 36 and then two cycles through the 72, making 252 presentations in all.

Sessions 2, 3, 4 - Training. Here the groups were differentiated according to block size as shown in Table I. Each session consisted of four random cycles through the block of anticipation learning trials, then four through the next block, etc, all random orders differing from one another. Each session ended with a test of that day's 72 in a new random order, making 288 + 72 = 360 trials per session.

Sessions 5 and 6 - tests of the 216. While not directly relevant to the model, these tests were included to provide data on retention over days, because only within-session retention was measured during Sessions 1-3. The fourth session occurred 24 hours after the third, and comprised one random cycle through all 216 items, each recall response being followed by reinforcement. The fifth session occurred one week later, and measured retention of items learned on or before their Session 4 reinforcement.

Results and Discussion

First we examine the test data from Sessions 1-3, since these are directly relevant to the model's predictions. Then we discuss the acquisition data from these sessions, followed by an analysis of retention based on results from Sessions 4 and 5.

Table 2 gives the mean and standard deviation of the

Table 2 here

frequencies of correct responses on Tests 1-3, i.e. the tests of 72 items after four reinforcements per item. According to the lower right-hand entry, the proportion correct averaged over all block sizes and subjects was .47. Note from the table that in five of the six groups the mean frequency correct under size 18 exceeded that under size 36. The same five groups averaged more correct with size 18 than with size 72. For sizes 18, 36, and 72, the mean frequency correct was 36.25, 33.40, and 32.15, respectively. Therefore the pairwise comparisons among within-group means, and also these three means across groups, reveal a slight tendency for size 18 to surpass size 72 in test frequency correct, with size 36 intermediate. But clearly the differences are not statistically significant. The model, then, receives no more than equivocal support. That is, the two above analyses uphold the prediction that the proportion correct should vary monotonically with block-size. Contradictory to the model, however, is the fact that only three of the six groups in Table 2 exhibit monotonicity among the three within-group means.[1] Further analysis indicates that even the conclusion of moderate support requires qualification. Evidently the order in which the block-sizes were presented influenced learning as much as did the size itself. This conclusion follows from Table 3, which summarizes individual subject analyses. By way of

Table 3 here

illustration, consider the first two rows of the table, which summarize the comparison between sizes 18 and 36. Within the three groups for whom size 18 preceded size 36, 9 subjects had more correct on the size 18 test but 11 had more correct on the size 36 test. Rows 1, 3, and 5 of the table show that when the smaller block size preceded the larger, test performance was virtually equivalent under the two sizes. Rows 2, 4, and 6 of the table show that the superiority of the smaller block size was limited to the groups for whom this size followed the larger size. A concise statement of the within-subject analysis would be: (1) of the 144 comparisons (48 Ss x 3 pairs of block sizes), the smaller size yielded more learning in 79, the larger size in 55, with the sizes equal in the remaining 10 cases; (2) for 80 of the 144 comparisons the temporally later block size surpassed the earlier, _independently_ of the size.

In passing, we mention a plausible hypothesis that is easily refuted by the Table 2 mean data. The notion is that the shift in block size from one set of items to the next, rather than absolute size, affects test performance. Contrary to this suggestion, groups 18-36-72 and 72-36-18 exhibited the same rank order of block sizes, despite the reversal in order of presenting the sizes.

How accurate is the theoretical prediction, considering Tables 2 and 3? Of course, the model has no provision for the "learning set" phenomenon that presumably underlies the order effect seen in Table 3. Before the prediction can be evaluated we must either (1) achieve a more or less statistical control by averaging across presentation orders (or subjects) or (2) compare two block sizes at the same stage in the experiment. Alternative (1) is the analysis already reported and based on Table 2. Alternative (2) entails looking at data from the first block-size a subject received. From Table 2, groups 18-36-72 and 18-72-36 yield a mean frequency of 34.13 correct under size 18. The other appropriate pairs of groups give 33.07 for size 36, followed by 32.38 for size 72. Therefore this analysis upholds the conclusion drawn from analysis (1); namely, that the mean performance under block-size 18 is superior to that under 36, which in turn surpasses that under size 72.

The standard deviations in Table 2 attest to the wide inter-subject differences in frequency of correct responses on Tests 1-3. It is curious that despite variations in means, schedules, and subjects, five of six groups displayed their highest standard deviation under size 36. Unfortunately, groups 18-36-72 and 18-72-36 yielded widely discrepant frequencies correct under size 18. It is because of such variability that we have relied on intragroup analyses to evaluate the block-size prediction. But there was considerable intrasubject variation as well. This fact emerges by computing rank-order correlations. Within each group, the eight subjects were rank-ordered three times, the orders being based on the Tests 1-3 frequencies correct under the three sizes. A statistic can be computed which has approximately a χ^2 distribution with, in our case, 8-1 = 7 degrees of freedom (Siegel, 1956). Listing the groups in the order given by Table 2, χ^2 = 13.33, 10.75, 14.11, 11.53, 13.03, and 12.30. (χ^2_{10} = 12.02 and $\chi^2_{.05}$ = 14.07). Thus on the whole the computed correlations are not statistically significant, pointing up a tendency for the within-group rank-order of subjects to vary from one block size to another.

Continuing the analysis of data from Sessions 1-3, we now turn from the tests to the acquisition phase which immediately preceded each test. The relevant data are the frequencies correct on each of the four cycles through the N items, where N is the block size. Table 4 presents the findings for size 18, while data from sizes 36 and 72 appear in Table 5. For a particular group, and cycle (1st, 2d, 3d, or 4th), each entry is a mean over the eight subjects' frequencies correct. The tables do not include cycle 1 for each set, because all entries would be zeros on this anticipation cycle prior to the first reinforcement.

Tables 4, 5 here

A number of qualitative conclusions could be drawn from Tables 4 and 5; a few observations suffice for our purposes. For one thing, there is only slight evidence of "learning-to-learn"; the last row of Table 4 shows modest improvement from the first to the last block of 18 items. Although the group mean proportions correct increase monotonically over cycles 2-4 within a set, some of these curves do not display the negative acceleration predicted from, e.g., a one-element or single-operator linear model. For example, Table 4 shows that group 36-18-72 manifested constant or increasing, rather than decreasing, gains within the first, third, and fourth sets of 18 items. Positive acceleration also obtained with size 72. In fact, a closer scrutiny of the last three columns of Table 5 reveals a striking interaction between the learning curve acceleration and the time when set 72 was presented. That is, performance under size 72 revealed a pronounced positive acceleration for the two groups who had size 72 as their first set, a less marked upswing for groups 18-72-36 and 36-72-18, and negative acceleration for groups 18-36-72 and 36-18-72. Numerically, the increment from cycle 3 to cycle 4, minus the increment from cycle 2 to cycle 3, was 2.3 and 2.0 for the groups who began with size 72, against 1.7 and 1.7 for the groups who had size 72 last. The suggestion arises that the former, positively accelerated curves, although starting from a lower frequency correct, would eventually overtake the negatively accelerated curve. Indeed, on size 72, cycle 2 and 3, groups 72-18-36 and 72-36-18 average fewer correct than groups 18-36-72 and 36-18-72, but the inequality is reversed on cycle 4. To the extent that it is valid, extrapolation suggests that if more cycles had been run the latter two groups would have fallen below 18-72-36 and 36-72-18 also.

This interaction between block schedule and learning curve acceleration, most notable with size 72, discourages a more detailed application of the one-element or linear models that underlie the block-size model. At the gross level of predicting the optimal size for mean performance, the model is supported, albeit only weakly. But a more elaborate formulation is needed to fit the details in Tables 4 and 5. We shall be content to infer from Table 5 some of the requisite properties for a more adequate model. Judging from the size 36 means for five of the six groups, the model must permit a mild learning-to-learn effect from one set of 36 to the next, regardless of the block schedule. Beyond this between-set effect the data cited in the last paragraph argue for a within-set mechanism depending on the particular block schedule. One possibility that might be developed is Norman's (1964) two-phase model, in which incremental learning begins after a random number of trials that yield no learning. However, this model has not proven particularly accurate in handling paired-associate data. Or some insight might be offered by long-term, short-term retention models that permit intertrial forgetting (Atkinson and Crothers, 1964; Calfee and Atkinson, 1965). Before such cursory speculations bloom into a model, it would be well to determine (i) whether the crossing of learning curves is genuine or represents random fluctuation about a common asymptote, and (ii) whether the shift from positive to negative acceleration would have occurred under a block schedule such as 72-72-72.

Sessions 4 and 5. These data (see the last two columns of Table 2) were collected more for their relevance to longer-term retention than for their immediate theoretical significance. Based on 216 items, the mean proportion correct over all groups was 64.27 on Test 4 and 58.04 on Test 5. Highest test frequencies correct were associated with the two groups that learned size 72 first. Columns 4 and 5 of Table 2 reveal some disparity between the rank-order of groups based on mean Test 1-3 performance and that based on Test 4 scores. Probably the reinforcements on Tests 1-3 are largely responsible for this reordering. No doubt the impressive degree of retention from Test 4 to Test 5 one week later is in large measure attributable to the occurrence of reinforcements on both tests.

Probing more deeply into the Test 4 data, we partitioned the 216 items according to the block-size under which they had been learned. Of course, this analysis is merely suggestive, because (i) block-size and block order are confounded, and (ii) the item may have been learned on, but not before, the test that terminated one of the three learning sessions, in which case the acquisition block-size is irrelevant. Table 6 gives

Table 6 here

the frequencies correct, broken down into acquisition block-size, and summed over the 8 subjects per group. The outcome, although internally consistent, is hardly surprising. Within each group, the block recency variable completely overshadows block-size in determining performance 24 hours after the third block-size. Nonetheless, the three column means reveal a small superiority for size 18. Presumably the latter outcome is due to more items being learned under size 18 rather than enhanced resistance to forgetting. Relevant here would be the response probability conditional on prior responses to that item, but the two above mentioned limitations of this analysis, and the small magnitude of the effect, render such a computation of little interest.

Summary.

This experiment continued the series of tests of the empirical adequacy of a model for predicting optimal block size. If there are X items and X = YZ then by "block size Y" we mean that the first set of Y items get all their reinforcements first, followed by the second set of Y, etc. and finally the Zth set of size Y. Optimality is defined in terms of performance on a test after the Zth set. The design formerly used was now modified by running the same subject under all three (18, 36, and 72 - item) block sizes, with six groups of subjects corresponding to the six permutations of the order of three sizes. With each size, 72 Russian - English pairs received four reinforcements each, followed by a test of the 72.

The order of block-sizes in terms of mean test frequency correct, from highest to lowest, was 18-36-72, an agreement with the monotonicity prediction of the model. However, differences among sizes were not statistically significant, and were overshadowed by the tendency for more correct responses on later than on earlier blocks, regardless of the block-size. Retention after a seven day interval depended more on the item's presentation recency than on the size of the block in which it appeared. Thus, the session (first, second, or third) in which a given item occurred was the major determinant both of within-session retention and retention one week later.

Footnotes

[1] Taking the two most proficient subjects in each group (based on total errors summed over Tests 1-3) and pooling over groups, the mean frequencies correct were 45.83, 46.33, and 42.58, for sizes 18, 36, and 72, respectively. On the other hand, the least proficient subjects yielded corresponding means of 27.42, 24.17, and 24.67. Roughly speaking, therefore, the shift to a slightly lower mean frequency correct came at a smaller block size for the poorer subjects than for the better subjects.

TABLE 1

	Counter-balanced Block Size Presentation Order For Sessions 2-4.*		
Group	2	3	4
18 A	18	36	72
18 B	18	72	36
36 A	36	18	72
36 B	36	72	18
72 A	72	18	36
72 B	72	36	18

*Sessions 1, 5, and 6 were identical for all groups. Session 1 - pretraining. Sessions 5 and 6 - test and retest, respectively, on the 216 items.

TABLE 2

Frequency of Correct Responses on Tests 1-3*

	Block Size					
Group	18	36	72	mean	test 4	test 5
18-36-72	40.4	34.1	32.4	35.6	66.6	59.4
	15.1	15.2	10.3	13.5	32.5	28.3
18-72-36	27.9	31.4	28.5	29.3	52.8	50.3
	9.3	12.2	10.3	10.6	20.2	20.3
36-18-72	35.3	33.8	32.1	33.7	59.4	51.4
	8.6	4.1	10.8	7.8	18.7	20.7
36-72-18	37.1	32.4	35.1	34.9	63.8	54.5
	6.4	7.9	7.1	7.1	12.9	15.6
72-18-36	35.6	34.9	35.1	35.2	70.9	65.4
	10.5	13.4	10.3	11.4	20.8	15.1
72-36-18	41.3	33.9	29.6	34.9	72.3	67.4
	7.9	10.5	7.2	8.6	18.4	27.1
mean	36.3	33.4	32.2	33.9	64.3	58.0
mean proportion	.50	.46	.45	.47	.30	.27

*Upper entry in each cell is the mean, and lower entry is the standard deviation.

TABLE 3

Pairwise Comparisons of Block-Sizes, Based on Individual Subject Frequencies on Tests 1-3

18 vs. 36:	18 > 36*	36 > 18
18 before 36	9	11
36 before 18	15	8
18 vs. 72:	18 > 72	72 > 18
18 before 72	12	10
72 before 18	16	7
36 vs. 72:	36 > 72	72 > 36
36 before 72	11	12
72 before 36	16	7
column sum	79	55

*For example, the first entry (9) means that of the subjects who received block size 18 in an earlier session than size 36, 9 made more correct responses on the size 18 test than on the size 36 test. The reason each row total is less than 24 is that some subjects made exactly as many errors with one block size as with another.

TABLE 4

Mean Frequency of Correct Responses on Training Trials, Block Size 18

	Sets of 18 items											
Group	1st			2nd			3rd			4th		
	2	3	4	2	3	4	2	3	4	2	3	4
18-36-72	3.5	5.8	8.8	4.8	8.4	10.6	4.6	8.9	11.9	5.1	9.4	12.4
18-72-36	2.4	4.5	6.6	3.6	6.1	9.8	4.9	7.6	8.8	4.6	8.0	11.3
36-18-72	4.1	7.8	11.5	5.3	8.5	11.4	4.1	6.6	9.4	3.8	6.2	9.3
36-72-18	3.6	6.3	9.1	3.8	7.9	11.4	6.4	9.9	12.9	5.0	8.9	10.9
72-18-36	4.1	8.0	11.1	4.6	8.1	10.6	4.0	7.6	10.3	2.8	6.8	10.3
72-36-18	3.0	7.1	9.5	4.3	9.1	11.9	5.9	11.0	14.6	5.5	9.8	12.1
mean	3.5	6.6	9.4	4.4	8.0	11.0	5.0	8.6	11.3	4.5	8.2	11.1

TABLE 5

Mean Frequency of Correct Responses on Sessions 1-3 Training Trials, Block Sizes 36 and 72

	Sets of 36						Set of 72		
Group	1st			2nd			1st		
	cycle no. 2	3	4	cycle no. 2	3	4	cycle no. 2	3	4
18-36-72	4.9	10.6	16.5	7.1	12.6	17.3	7.6	18.1	24.8
18-72-36	4.9	8.4	11.9	6.6	12.5	17.4	6.6	13.1	21.3
36-18-72	4.1	9.9	14.8	6.6	13.3	18.1	8.9	18.1	25.1
36-72-18	4.3	8.5	14.8	6.9	12.4	19.1	8.0	16.9	27.5
72-18-36	5.1	11.5	16.5	6.4	10.5	16.1	8.0	16.9	28.1
72-36-18	5.4	11.1	16.0	6.8	13.0	17.5	7.8	15.3	24.8
mean	4.8	10.0	15.1	6.7	12.4	17.6	7.8	16.4	25.3

TABLE 6

Group Total Frequencies Correct on Test 4,
Classed According to Acquisition Block-Size

Group	Block Size			
	18	36	72	∑
18-36-72	17.6	22.6	26.4	66.6
18-72-36	11.8	25.9	15.1	52.8
36-18-72	20.4	13.6	25.4	59.4
36-72-18	30.3	13.6	19.9	63.8
72-18-36	22.1	27.8	21.0	70.9
72-36-18	34.3	22.0	16.0	72.3
∑	136.4	125.5	123.8	385.6
mean	22.7	20.9	20.6	64.3

TABLE 7***

Median Frequencies of Correct Responses

Nouns and Pronouns						
Pronounciability	High assoc. val. (AV)			Low assoc. val. (av)		
	1	2	3	1	2	3
High (P)	(8) 112.5	(16) 67		(12) 51	(41) 36	(15) 36
Low (p)	*	*		(7) 5	(11) 17	(11) 20

Other Form-Classes						
	High assoc. val (AV)	Low assoc. val. (av)				
	adj.	adj.		verb		adv.
	2	2	3	2	3	2
High (P)	(12) 75	(15) 30	(19) 28	(21) 12	(11) 8	(9) 11
Low (p)	*	(11) 18	*	(7) 9	*	(5) 5

*** Cells denoted by an asterisk (*) or omitted from the table contained fewer than five words. Numbers in parenthesis denote the number of words entering into the computation of the cell median.

TABLE 8
Mean Proportions Correct

		response	
		easy	hard
Stimulus	easy	.454*	.299**
	hard	.118**	.015*

*Data from Suppes-Crothers exp. Days 1-10.

**Data from Rodgers exp, trials 1-6.

Optimal Presentation Orders for Items from Different Categories

This generalization of Suppes' model was reported as Institute for Mathematical Studies in the Social Sciences, Tech. Rep. No. 71, (Crothers, 1965), copies of which have been submitted earlier. Here we include an abstract of an expanded draft of the report:

Other researchers have published considerable experimentation on the problem of optimal presentation order for items from diverse categories. Specifically, we assume that the number of categories is C, with E exemplars each, making a total of EC items to be learned. The question is whether learning is better under maximal clustering of similar items (high contiguity) or under maximal separation (low contiguity). Three processes are assumed, learning of an item when it is presented, positive transfer when another member of the same category is presented, and forgetting when a member of a different category occurs. Thus by the transfer mechanism we relax Suppes' assumption of independent items. Two sets of derivations are carried out for each level of contiguity, the sets effectively distinguish between reinforced and nonreinforced test cycles. In the reinforced case, the analytic solution is that high contiguity should be the more efficient. In the unreinforced case there is no parameter-free solution, but numerical explorations suggest that performance depends more on the number of concepts than on the number of exemplars. Also, the more rapid the learning, the more likely it is that low contiguity will be superior. A review of published experiments revealed strengths and weaknesses of the model, and suggested that paired-associate categories defined on the basis of stimulus similarity may have different contiguity phenomena than categories which are concepts.

Linguistic Properties of Vocabulary Items and Rate of Learning

Here we report a further analysis of data from a vocabulary experiment described in last year's final report (OE 3-14-020, p. 5). Basically, subjects learned 300 Russian - English pairs, with auditory stimuli and experimenter-paced recall. Analyses formerly reported indicated the superiority of block-size 100 over size 300. The size 300 data represents a list that is large enough to provide an unusual opportunity for determining item difficulty as a function of its structure, and the gist of this subsequent analysis is our concern here. The main finding that will emerge is that the sources of difficulty accord with what would be expected from well-known studies with nonsense syllables (e.g., Underwood and Schulz, 1960).

In terms of the auditory Russian stimulus, the analyses parceled out the effects of the three most cogent variables: association value, pronounciability, and word length in syllables. Association values were roughly calibrated by pronouncing the words to three additional subjects and noting whether or not they could say any English associate. Pronounciability was rated on the basis of consonant clusters and adjacent vowels. Omitting the technical details of the qualitative scaling, each pair was classified thus: (1) Russian stimulus word; high (AV) or low (av) association value, high (P) or low (p) pronounciability, and number of syllables (1, 2, or 3). (2) English response; (1) form-class (noun-pronoun, adjective, verb, adverb). A combination of one value on each dimension is a cell.

Table 7 gives the median frequencies correct for each cell, based on the group 300 data summed over subjects and over days 1-10. The upper half of the table gives medians for the noun-pronoun class. For other form-classes, only the cells based on five or more items are displayed (in the lower half of the table). Since the data are not suitable for a factor-analysis, Mann-Whitney

Table 7 here

U tests were performed to ascertain the effect of each of the four dimensions.

Association value yielded a statistically significant effect. A typical result is that with one-syllable P noun-pronouns, $U = 10$ ($p < .01$). Pronounciability was also a powerful factor, as attested by the comparison between the two levels for the one-syllable, low av., noun-pronoun items ($U = 16$, $p < .025$). Fixing association value, pronounciability, and number of syllables, another analysis showed that verbs and adverbs were significantly more difficult than the other classes. On the other hand, the effect of the number of syllables was slight. The three dimensions which yielded significant effects demonstrate that well-known results in the literature on nonsense syllables carry over when the stimuli are Russian.

Rodgers, in a further study on this project, extended the findings by showing that properties of the Russian stimulus (association value and pronounciability) predominate over properties of the English response (form-class) in determining item difficulty. He took the 25 pairs with the most correct responses in group 300, plus the 25 with the fewest correct, and constructed pseudopairs as follows. Each stimulus member from the easiest 25 was paired with a response member from the hardest 25; likewise, each stimulus from the hardest 25 was paired with a response member from the easiest 25. If stimulus properties control difficulty, then a pseudopair with a difficult stimulus and an easy response should be relatively difficult - more so than a pseudopair with a formerly easy stimulus and a formerly difficult response. As Table 8 shows, this is exactly what happened. A plausible conjecture is that the unfamiliar member of the pair is dominant in determining difficulty; i.e., that even if the Russian word were the *response* member, it would still control the wrror rate.

Table 8 here

An Experiment on Learning Russian Syntax

The grammar of a second language poses one of the most formidable obstacles to mastery of the tongue. It is on the levels of syntax and semantics, rather than morphology, that the psychological problems of greatest theoretical significance arise. Inflection learning, as investigated in the preceding chapter, is actually a secondary embellishment of the grammar substrate. A speaker with adequate command of syntax and semantics could easily be understood despite deficiencies in his morphology usage, whereas little would be communicated if the subject could not compose meaningful sequences of words. Therefore it is hardly astonishing that the investigation of syntax learning is being increasingly pursued by psychologists.

Some insight into second-language learning can be gained from studies of first-language learning, although the latter is only slightly better understood. Perhaps the most outstanding aspect of the child's native language development is that he learns almost exclusively by hearing grammatical utterances, not by being told explicit rules of speech. Chomsky (1959) has argued that an essential prerequisite to a theory of language learning is a comprehensive specification of exactly what is learned. Examing syntactic structures independently of their semantic content, he has ably propounded the thesis that first language competence implies internalization of a highly formalized transformational grammar. Little is known about the mechanism responsible for the learned components of such a grammar. Nonetheless, his isolation of syntactic factors highlights the need for experimental analyses of syntax learning with minimal semantic information available to the subject.

Unquestionably, the skills of the native language shape the subsequent learning of new languages. But exactly what is transferred? We suggest that from the outset the novice has at his disposal habits of inference with which to build his provisional grammar of the second language. Not only does the student possess devices for learning grammar divorced from semantics, but many instructional methods encourage him to do so. Grammar rules stated explicitly or in tabular form are a familiar part of many courses. Although these reasons are hardly unassailable, they do furnish fairly cogent grounds for maintaining that experimental isolation of syntactic factors is at least as relevant for second - as for first - language learning.

Broadly speaking, the present experiment by Crothers and Rodgers involved Russian grammar acquisition and transfer as a function of the training conditions. The subjects first heard a series of grammatical Russian sentences and then had a variety of transfer tests. All independent variables were manipulated as between group treatments during the initial acquisition stage. One major objective was to see how syntax learning is affected by the concomitant presence of semantic information provided by English translations. Secondly, the presentation order variable was intensively explored. Lacking any formally generated predictions, it was conjectured on the basis of well-known evidence from other research areas that some orders, being more compatible with the subject's putative memory and information processing skills, would lead to more efficient test performance than others.

To be specific, the subjects received sentences constructed from small vocabularies and essentially had to learn what conditions an utterance must fulfill in order to be grammatical. They had to be able to handle a number of words in a variety of grammatical paradigms; namely, the declarative, negative, question, negative question, and passive. Presentation order manipulations were motivated by the expectation that two major sources of difficulty would be vocabulary variation and paradigm variation. One condition, hereafter called the permutation order, minimized vocabulary variation over successive trials while maximizing paradigmatic variation. Conversely, in the string constituent condition the paradigm remained constant over adjacent utterances but all words were replaced by others. A third condition, which will be called substitution-in-frame, was analogous to a frequently used educational method. As in the string constituent order, the paradigm remained the same from one trial to the next; the difference was that now part of the sentence was kept identical over several successive trials, while systematically substituting a different word in the other slot. Finally, a random order control condition was run.

Grammar learning could quite conceivably depend not only on syntax but on the phonological features of the utterance as well. Obviously some phoneme sequences are perceived and spoken more readily than others. Therefore, holding fixed the set of syntactic structures, a third variable was the particular vocabulary. Finally, transfer was measured in two ways. One technique was to run the subject with one vocabulary set and presentation order, then to repeat the experiment in a second session but with a different vocabulary and order. More importantly, intra-session transfer was gauged by including a wide range of structures on the tests. The first test was addressed to the question "what cues does the subject use in deciding whether or not an utterance is grammatical?" On this test he heard both grammatical and ungrammatical strings that deviated from the training sentences in highly specific features of syntax, morphology, intonation, and vocabulary. His judgments as to which of these are grammatical offer evidence concerning his functional cues. The final test explored what speech production skills had been built up depending on the absence vs. presence of semantics, presentation order, and vocabulary conditions.

Method

Design and experimental groups. The experiment consisted of three phases. First there was an acquisition phase, in which the subject heard a sequence of 160 grammatical Russian sentences, responding to each by imitating it aloud. Next occurred a classification test, on which the subject heard a series of 67 utterances and indicated after each whether or not he thought it was grammatical. Finally, there was a controlled speech production test in which the subject was asked to produce grammatical sentences.

All independent variables - presence vs. absence of semantic information, presentation order, and vocabulary - were manipulated as between - group treatments during the acquisition phase. Each subject went through all three phases with one presentation order and vocabulary set on his first day, then a different order and vocabulary set on his second day. Denoting the permutation, string-constituent, substitution-in-frame, and random order conditions as P, SC, SF, and R respectively,and using the subscripts I and II to identify the vocabulary sets, the first six groups are $P_I - R_{II}$, $SC_I - R_{II}$, $SF_I - R_{II}$, and $R_{II} - P_I$, $R_{II} - SC_I$, $R_{II} - SF_I$. That is, group $P_I - R_{II}$ had condition P_I on day 1 and condition R_{II} on day 2, with similar notation for the other groups. The vocabulary control group is designated as $SF_{II} - R_I$. The single group who received English translations, but only during learning, participated for one day and is called group $SF_{sem.I}$.

Each group contained 10 Stanford University students who were paid $1.50 an hour. None had prior knowledge of any Slavic languages and none had participated in any of our previous experiments. Assignment of subjects to groups was randomly determined by the chronological order of volunteering for the experiment, except that to save time the subjects were run in pairs, both members of the pair being under the same condition. Pair members were in separate rooms and did not interact in any way. The two sessions were separated by a twenty-four hour period.

Apparatus. The acquisition, classification test, and production test items and instructions were pre-recorded on an Ampex 601 tape deck, using normal speaking rate and inflection. The output apparatus consisted of the Ampex 601 tape deck coupled to an Ampex 620 output amplifier. The output signals were sent to a pair of Roberts 54-55 headsets at the two subject positions. Volume levels at the headsets were individually adjustable. Each of the two positions was equipped with a headset, microphone, and (for the classification test) a 67-blank answer sheet.

The subject's speech input was picked up on a high-impedance microphone at each position. (At position 1 a Sony Model A-7 was used and at position 2 an Electro-Voice Model 623 was used.) These input signals were recorded simultaneously on tracks 1 and 3 of a Crown Series SS800 four-track stereo recorder. The Ampex 620 output to the subject positions was simultaneously recorded on tracks 2 and 4 of the Crown Recorder. Thus all signals to and from Position 1 were recorded on one tape half-track, consisting of channels 1 and 2, and all signals to and from Position 2 were recorded on the other tape half-track, consisting of channels 3 and 4. This parallel recording of input and output facilitated monitoring responses and aided subsequent evaluation of the recorded speech.

Procedure. Stimuli in all stages of the experiment were in the auditory modality only. Each stage was prefaced by appropriate instructions (see Appendix A). The gist of the acquisition instructions was that the subject would hear grammatical sentences and, after hearing each, he was to imitate it aloud. He was warned that later his tasks would be to decide whether or not utterances were grammatical, as well as to produce grammatical sentences. No information was provided concerning Russian grammar. Then the 160 grammatical sentences were presented, each separated by a six second pause to allow time for imitation. Group $SF_{sem.I}$ heard the sentence translation after each imitation. Altogether this phase required about 30 minutes, and only slightly longer in group $SF_{sem.I}$.

Following a brief interval for changing tape reels, the classification test ensued. After hearing an utterance the subject had six seconds in which to write a + or a 0 in the blank on his answer sheet, according to whether he thought the utterance was or was not grammatical. The instructions that initiated this phase illustrated English strings that were ungrammatical because of word order, morphology, or word omissions. This phase required about 16 minutes.

The production test was a word substitution task, consisting of a grammatical, familiar paradigm sentence followed by a single familiar or new word. The subject was to substitute this word for one of the words in the original sentence, saying the new sentence aloud. The items used in each phase are described in the next section. No group received translations on either the classification test or the production test.

Materials and Presentation Orders

Acquisition. All sentences were positive (i.e., grammatical) instances constructed from one or the other of two 12-word Russian vocabularies. Each set consisted of twelve Russian words, eight nouns and four verbs (V), plus the particles ЛИ (indicating the interrogative) and the negation НЕ ("not"). Appendix B lists the words and their translations. The verbs were present-tense transitive (set I) or past-tense transitive (set II). The reason for changing the tense was so that the suffix distinguishing verbs from other parts of speech would differ from one set to the other. The eight nouns were divided according to their function in the declarative sentences; four masculine animate nouns were inflected as subjects (S) and four masculine inanimate nouns as direct objects (O). Our notation for the inflectional suffix is a superscript to the form-class symbol; e.g., $S^{nom.}$ is a subject noun stem with the nominative case inflection. By choosing all four S words from one gender and all four O words from a different gender, we were able to fulfill an important requirement - that each word contain a phonological cue sufficient to determine its form class. For example, in set I all $S^{nom.}$ words ended in-/ИК/, all verbs in -/ИТ/, and all $O^{acc.}$ in -/ИН/.

Another regularity internal to each set of four was the stress pattern; each word contained two syllables and all members of the same form-class were stressed on the same syllable. Since there were three form-classes and only two syllables, obviously the redundant stress cue was less useful than the suffix as a signal or form-class. At the sentence level, this stress regularity ensures that all sentences of the same paradigm will have identical word intonation patterns. For example, we see from Appendix B that in all SVO (i.e., declarative) sentences the first, first, and second syllables are stressed on the first, second, and third words, respectively. Because this 1-1-2 pattern does not characterize all paradigms, it should facilitate discriminations among them. A final criterion imposed on the selection of a 12-word set was that none of the $4^3 = 64$ possible SVO sentences be nonsense semantically. Altogether, these restrictions on suffix, number of syllables, stress pattern, and meaning drastically reduced the population of words eligible for inclusion in a set. Lastly, in order to study transfer from one vocabulary to another, no word (other than the particles ЛИ and НЕ) or suffix appeared in both sets.

After thus selecting the vocabulary, the sentences were generated in two steps. The manner of generating sentences from vocabulary was identical for both sets and will be illustrated with set I. First all $4^3 = 64$ ordered SVO triples of declarative (D) sentences were formed. Then each was transformed into negative (N), question (Q), negative question (NQ), and passive (P) paradigms, making 64 x 5 = 320 sentences in all. Table 1 summarizes the syntactic and semantic properties of these paradigms. Note that they can differ from one another in three respects:

Table 1 here

the particular stems and particles present present, the word order (syntax) and the inflection (morphology). One striking characteristic is the inversion of subject and verb in going from D or N to the other paradigms. Another feature concerns the particles; when НЕ is present it must immediately precede the verb, and when ЛИ is present it must immediately follow the verb. Note also that in the passive the imperfective verb ends in -СЯ, with the actor in the instrumental case. Numerous other properties can be seen from the table. To complete the paradigm description, we must specify what phoneme suffixes are represented by notation such as S^{nom}. Table 2 provides this information. Of the 320 sentences generated in the manner

Table 2 here

described only 160 were actually presented during learning. These consisted of the same 32 SVO triples in all five paradigms. These 32 were selected from the 64 so that each S, V, and O appeared equally often; i.e. in $\frac{32}{4} = 8$ of the triples, and each doublet (e.g., S and O) appeared with two of the four possible members of the third form-class. The reason for training on only a systematically selected half of the 320 sentences was to permit later introduction of new, unpracticed triples on the classification test.

Presentation Order. Table 3 shows schematically how the conditions differed in the order of presentation of the 160 sentences during acquisition. The subscript indicates a particular word and can range from 1 to 4; the suffix superscript and particles, being irrelevant here, are omitted. Condition R (random) was a random series of the

Table 3 here

160 items. Under condition P (permutation) each nonoverlapping block of five presentations had the same SVO triple, in the order D, N, Q, NQ, P. The next block changed to a different S, V, and O and repeated the paradigm order, so each of 32 triples occupied one five-trial block. Thus from one presentation to the next there was maximal paradigm variation and minimal vocabulary variation. The opposite condition, successive-trial variation in vocabulary rather than paradigm, characterized condition SC (string constituent), so named because the type of string (sentence) was held constant over a series of 32 trials while vocabulary items in consecutive string presentations were never repeated. The first 32 trials involved declaratives, the next 32

negatives, etc., conforming to the D, N, Q, NQ, P order from one set of 32 trials to the next rather than from one trial to the next. Within a block of 32, each SVO triple occurred once, the order being random except for the restriction that successive presentations did not repeat vocabulary items. Condition SF resembled condition SC in again devoting the first 32 trials to declaratives, the next 32 to negatives, etc. But within each block of 32 the SF sequence adhered to a rigid ordering. The general idea was that a particular subject and verb frame remained unchanged while the object was varied over a sequence of successive trials. One of the four masc. anim. nouns was the subject in the first eight sentences, another in the second eight, etc. Further, the verb changed only on odd-numbered trials. On each trial a different one of the four objects was substituted in a frame consisting of a subject that changed every eighth trial and a verb that changed every other trial. That identical substitution order was used for all five blocks of 32 trials. To avoid ambiguity we mention that the term "block" does not mean a temporal segregation of items; the 160 presentations followed without interruption and at a uniform interitem interval.

Classification Test. The classification test was identical for conditions P_I, SC_I, and SF_I. That for R_{II}, SC_{II}, and SF_{II} was constructed according to the same principles but with vocabulary set II instead of I. The sixty-seven items were presented in random order. To control for possible effects of test sequence, approximately one-third of the subjects in each condition had each of three random test presentation orders. Upon hearing each utterance the subject judged whether or not it was a grammatical sentence. No information was given following the response.

The 29 grammatical (G), and 38 ungrammatical ($\bar{G}$) strings were a heterogeneous set constructed so as to probe how structural factors such as improper syntax, improper morphology, improper intonation, unfamiliar words, and unfamiliar paradigms influence judgments of grammaticality. By ascertaining what sort of deviations from the acquisition sentences were rejected as ungrammatical, we sought to determine what linguistic properties of the utterance governed the binary decision. To control for possible word preferences, each of the twelve words occurred almost equally often in G and in $\bar{G}$ utterances.

Table 4 here

Table 4 summarizes details of the test item construction. To permit compact descriptions we introduce the following notation: when the symbols, S, V, and O appear without subscripts or superscripts they denote the declarative inflection $S^{nom.}$, V, $O^{acc.}$. Regarding the improper syntax items in part I of the table, Q (V, ЛИ) denotes a question rendered ungrammatical by transposing V and ЛИ , and similarly for N (V, HE). Also, P(O + CЯ) denotes a passive distorted by suffixing -CЯ to O instead of to V. In part II of the table, a bar above a symbol means the word was stressed on the normally unstressed syllable. In part II, D (-V) means a declarative with the verb omitted, and likewise for the other omission items. Morphology items are written with stems in upper case and inflections in lower-case subscripts. For example, $S_vV_oO_s$ means a grammatical SVO stem sequence, with S, V, and O artifically given verb, object, and subject inflections, respectively, thereby producing pseudo-words.

Among G items in part I of the table, the variables are the novelty (familiarity) and paradigm (D, N, Q, NQ, P) of the utterance. The most familiar items are the familiar SVO triples - a subset of the 32 triples presented during training. The category entitled "familiar words" designates a subset of the other 32 triples, i.e. those not formerly presented. All items were familiar triples (deliberate distortions of morphology, or stress in $\bar{G}$ items do not affect this classification) except those in rows 2 and 3 of part I of the table. The category "one new word" differs from all other item types in that each utterance contained one word not previously heard during learning. The new words could be subjects, verbs, or objects, and are listed in Appendix B. The category "new paradigm" denotes familiar words arranged in paradigms the subject had never experienced, such as negative passive question (NPQ). The $\bar{G}$ items in part I represent cases where the syntax is incorrect. Here the new paradigms were excluded, leaving the other three degrees of novelty. Taking the six possible permutations of the elements S, V, and O, three of these (VOS, OSV, and SOV) do not appear in any paradigm and are therefore ungrammatical. In addition, the Q(V, ЛИ) and N(V, HE) transpositions were included to see if the subject had learned that ЛИ can only occur immediately after the verb, and HE only immediately before the verb.

Part II describes the items included to assess the role of word stress in judgments of grammaticality. Six anomalous intonation items had correct SVO syntax and inflections but in one or more of the two-syllable words the speaker deliberately stressed the wrong syllable. For example, in set I the proper SVO sentence intonation sequence was S- first syllable, V- first syllable, O- second syllable, and this 1-1-2 pattern was violated by these items in the manner shown in the table. Although these items (in Table 4) are linguistically ungrammatical the subject's interpretation of "grammaticality" as regards intonational features is somewhat unclear, so these data will be analyzed separately. Contrasting with these items were the items entitled "SVO stress", which preserved the 1-1-2 intonational pattern (set I) but had incorrect syntax.

A typical item here was $V\bar{O}\bar{S}$; the bars denote that when the correct stress pattern is 1-1-2 the ungrammatical VOS order implies that O and S are stressed on the wrong syllable. In short, the disparity between the anomalous stress and SVO stress $\bar{G}$ items was intended to pit word order patterns against intonational patterns and to see which was dominant in determining judgments of grammaticality.

Quite obviously, factors other than syntax and intonation are often responsible for an utterance being ungrammatical. Two such possible factors, the omission of morphemes and improper morphology, are represented by items in part III of the table. "Omissions" were utterances with elements such as НЕ and ЛИ intentionally omitted by the experimenter, whereas in the previously described improper syntax items these elements appeared at a wrong position in the string. Of the many possible sorts of morphology errors, two are relevant for our purposes: use of an inflection that would be correct for a different paradigm (e.g. $O^{gen.}$ instead of $O^{acc.}$ in the declarative), and use of an inflection that belongs to a totally different form-class (e.g., a verb suffix on a noun stem). The latter error is the more flagrant and was represented in our $\bar{G}$ morphology items as follows. We inserted a "syntactic" condition, partitioning the incorrect morphology items according to whether it was the suffix sequence or the stem sequence that is ungrammatical. In the above notation, the stem sequence is ungrammatical in a string such as $V_sO_vS_o$, whereas only the suffix sequence is ungrammatical in type $S_vV_oO_s$.[1] The rationale for these morphology items was to ascertain whether the subject would make either of two erroneous inferences: that grammaticality depends only on the stem sequence, or that it depends only on the suffix sequence.

Production test. This test ended each session. Each item was a sentence followed by an isolated word. The subject was to decide which word in the sentence could be replaced by the isolated word, indicating his decision by saying the new sentence aloud. There were 45 items in random order, nine in each of the five paradigms. Of the nine, three required substitution of a new subject, three a new verb, and three a new object. Every to-be-inserted word was given with inflection correct for its role in the sentence. Note that the relation between a sentence and the correct new sentence is like that between successive sentences in the SF, but not SC, P, or R condition.

Results and Discussion

Classification Test.

Preliminary analysis revealed no differences in performance under the three different test random orders, so data from all subjects in a group were combined. Also, to test for learning effects over the series of unreinforced test trials, data from pairs of syntactically equivalent items were analyzed according to which member of the pair occurred earlier on the test. Such pairs included familiar sentence G items, new paradigms, and anomalous intonation items. There was no appreciable change in proportion correct from the earlier to the later members of the pairs. Lastly, monitoring of responses during acquisition indicated generally good imitation by the end of acquisition and no noticable intergroup differences. From these preliminary computations we conclude that any effects revealed on the classification test can safely be attributed to the acquisition conditions per se, not to differences in imitative skill or learning on the test.

Table 5 presents the data most directly relevant to the effects of semantics, vocabulary, and presentation order. Each row represents a category of items, each column a group of subjects, and each entry a Day 1 error frequency summed over all subjects in the group. Note that the four item classes in Table 5 are aggregates over the subclasses in Table 4 (e.g., the various types of ungrammatical items are pooled). This pooling was employed to produce sufficient observations per cell; data on the finer partition of the items appear in the next section.

Presence vs. Absence of Semantics. The pertinent comparison is between the $SF_{sem\ I}$ and SF_I columns of Table 5. Surprisingly, the former group made more errors than the latter on the grammatical (G) items but fewer errors on the ungrammatical ($\bar{G}$) items. For example, by converting from frequencies to proportions we see that the error proportions on the $\bar{G}$ items were $\frac{72}{320} = .225$ in group $SF_{sem\ I}$ against $\frac{112}{320} = .35$ in group SF_I. Mann-Whitney U-tests were performed using the individual subject error frequencies. For the familiar paradigm G items, $U = 19$ ($p = .02$), and for the $\bar{G}$ items $U = 17$, ($p < .02$). Hence presence of semantics significantly retarded performance on the former items, while facilitating it on the latter. These findings with the familiar paradigm G items and with all $\bar{G}$ items cannot be imputed simply to an indiscriminate tendency for $\bar{G}$ responses by group $SF_{sem\ I}$, because this group yielded fewer $\bar{G}$ responses on anomalous intonation items than did group SF_I.

A more promising alternative explanation is that subjects in group $SF_{sem\ I}$ judged grammaticality partly on the basis of whether or not they could translate the test stimulus utterance. This hypothesis would account for the fact that $SF_{sem\ I}$ subjects, more than SF_I subjects, tended to call the anomalous intonation strings grammatical; ($U = 26$, $.05 < p < .10$); the string was otherwise correct and intonation does not interfere with the translation. A second implication of this hypothesis is that group $SF_{sem\ I}$ would classify as ungrammatical utterances that contained a new word, because that word had never been translated. Exactly this happened on the familiar paradigm G items. According to Table 5, group $SF_{sem\ I}$ made 11 more errors than group SF_I on these items. More detailed analysis showed that this inferior performance by the group with semantics was limited chiefly to one subclass of familiar paradigm G items; namely, the ones that contained a new, untranslated word. By the same token group $SF_{sem\ I}$ should excel on the $\bar{G}$ new word items, but here the data are equivocal. What, then, is our conclusion about how the availability of English translations during learning affects subsequent judgments of grammaticality? It appears that exposure to English equivalents somewhat induces the subject to classify a string according to whether or not he can translate it. This strategy has the advantage of facilitating the detection of ungrammatical strings, but the disadvantage that a novel grammatical utterance (one with a new word, or a new paradigm), being imperfectly translatable, is erroneously judged ungrammatical.

Vocabulary Set I vs. Set II. Here the relevant comparison is between the SF_I and SF_{II} columns of Table 5. Strictly speaking, our conclusions are restricted to the

SF presentation order and to the case where no concomitant translations appeared during learning. Somewhat unexpected was the finding that on familiar paradigm G items subjects with vocabulary II made more than twice as many errors as subjects with vocabulary I. Additional analysis showed that this disparity occurred uniformly in the three subclasses of the familiar paradigm G items. The superiority of SF_I to SF_{II} is statistically significant; $U = 20$ and $.02 < p < .05$. No difference emerged on $\bar{G}$ items. On the intonation items set II yielded $\frac{10}{60} = 16.7$ percent more judgments $\bar{G}$ than did set I. If we adhere to the defensible position that anomalous intonation items are in fact ungrammatical, then the data from these and from the familiar paradigm G items indicate that subjects having vocabulary I were more accurate than ones having vocabulary II. It is noteworthy that except on the $\bar{G}$ items the effect on error frequency of the particular vocabulary was as large as the effect of semantics.

Presentation Order: SF_{II} vs. R_{II}. These results refer to groups who did not have English translations. Differences attributable to presentation order are very slight in the last two columns of Table 5. There is a hint that on G items SF surpasses R when the paradigm is familiar, whereas the order of groups is reversed when the paradigm is novel. But no formal statistical test is needed to establish that with vocabulary II test performance is essentially the same after the SF as after the R training order.

More light is shed on the role or presentation order when we consider jointly the nonrandom conditions P_I, SC_I, and SF_I. As we shall indicate presently, there appears a slight but consistent tendency for more correct judgments under SF than under P. Taken in conjunction with negligible difference between conditions SF_{II} and R_{II}, this finding suggests an interaction between vocabulary set and presentation order, the latter variable affecting performance only with set I. Table 6 gives the Day 1 error frequencies for each of the nonrandom conditions which had set I vocabulary.

Table 6 here

First we focus on the Day 1 data. Pairwise comparisons among the groups support the tentative conclusion that condition SF led to more accurate classifications than condition P, with the relative standing of condition SC being equivocal. The rationale for the asserted superiority of SF to P is that the former yielded slightly fewer errors than the latter on the new paradigm G items, and more $\bar{G}$ responses on the anomalous intonation items. Responses $\bar{G}$ on the intonation items should be regarded as correct and occurred significantly more often with SF_I than with P_I ($U = 19$, $p = .02$). The two groups were equivalent on the remaining Day 1 items. Although most of these intergroup differences are admittedly small, further indication that SF was more efficient than

P comes from the Day 2 data. This time the margin is largely restricted to the $\bar{G}$ items, where the proportion of errors under condition P exceeds that under SF by about .06. This Day 2 result cannot be dismissed as a subject selection bias, because on Day 1 group R_{II}-P_I yielded fewer, not more, errors than group R_{II}-SF_I (excluding the intonation items, the Day 1 means error frequencies were 19.0 and 20.2 respectively). Thus data from four groups hint that condition SF was more effective than condition P. Further analysis revealed that the effect was not strong enough to persist when the groups were shifted to condition R; groups P_I-R_{II} and SF_I-R_{II} performed equally well on the Day 2 classification test.

Despite the fact that the relative accuracy of conditions SC and SF is not clear from Table 6, it does appear that the contrast between SC and P is consistent across the two days and four relevant groups. That is, by comparison with SC, P yielded fewer errors on new paradigm G items, more errors on $\bar{G}$ items, and fewer $\bar{G}$ responses on the intonation items. By contrast with P, SC exhibited a striking pattern of more $\bar{G}$ responses to the "novel" items, regardless of whether this response was correct. This conclusion was established by calculating the total $\bar{G}$ responses for each of the four relevant item types: the G and $\bar{G}$ items that contained a new word, the new paradigm items, and the intonation items. With each type, SC produced more $\bar{G}$ responses than P, based on Day 1 data from groups SC_I-R_{II}, P_I-R_{II}, and Day 2 data from groups R_{II}-SC_I and R_{II}-P_I. None of the other group comparisons (R_{II} vs. SF_{II}, SF_I vs. SF_{II}, etc.) disclosed novelty effects strong enough to hold over both days and all four novel item types.

Transfer from Day 1 to Day 2. Intuitively, it is debatable whether or not to expect transfer from one presentation order and vocabulary on Day to a totally different order and vocabulary on Day 2. The results concerning this particular index of transfer appear in Table 7; for the analysis it suffices to lump all but the intonation items and compute a mean error total per subject based on 61 items.[2] No appreciable transfer occurred from R_{II} on Day 1 to any of the nonrandom Set I conditions on Day 2. Similarly, no improvement occurred from a nonrandom set I condition on Day 1 to R_{II} on Day 2. Indeed, even if there were no genuine improvement over days, the confounding present in our comparison might reveal an effect of vocabulary set or presentation order. But the decisive fact in Table 7 is that all of the frequencies are remarkably similar; this strengthens our main conclusion that the effects of the vocabulary and order variables are limited to a minority of specific item classes. At this point it is perhaps superfluous to note that another index of transfer, as illustrated by, say, the contrast between P_I in groups P_I-R_{II} and R_{II}-P_I, confirms

Table 7 here

the finding of negligible transfer. Supplementary tabulations, which consisted in pairing each subject's Day 1 and Day 2 error frequencies, corroborate this conclusion. A less important but interesting outcome is that for only one of the seven groups in Table 7 does the error range on Day 1 exceed that on Day 2.

Linguistic Correlates of Classificatory Responses. Now that we have discussed the principal intergroup comparisons, let us combine data across groups and ask how judgments depended on the linguistic properties of the utterance. The goal of this analysis is to determine what structural features of the utterance were used by the subjects as cues to grammaticality. Presumably, the most prominent cues would be derived from the syntax, morphology, intonation pattern, and familiarity of the string; the following analyses sought to assess the relative degree to which each of these features determined the binary judgments. The major findings will be reported here with the more detailed breakdown deferred until the next section.

First consider the items that were actually ungrammatical but consisted of familiar stems and suffixes. Recall from Table 4 that there were three such item categories: (1) omissions - utterances in which one of the elements verb, НЕ, ЛИ , or -СЯ had been omitted from an otherwise grammatical sentence; (2) improper syntax - words were transposed to produce an ungrammatical order, but each word has a proper inflectional suffix; (3) improper morphology, produced by attaching, say, a noun suffix to a verb stem. Table 8 shows the proportion of times that utterances of each type were incorrectly judged to be grammatical. The rank order from most to least difficult is: omissions, improper syntax, improper morphology. These differences from one category to another in proportion correct are strikingly large compared to the intergroup differences reported in the preceding section. For example, averaging over the groups without semantics in Table 8, the mean proportion of errors on the omissions items exceeds that on the morphology items by at least .40. It is also worth noting that for the seven groups without semantics the error proportion on the omission items exceeded by .12 or more the expected chance proportion of .50. Nonetheless, the contrast between rows 1 and 4 of Table 8 demonstrates that omissions were frequently detected; because the row 4 proportion of G responses ranged from .81 to .90 among familiar items that were in fact grammatical, but dropped to a range of .50 to .65 for omissions. Apparently the introduction of semantics was particularly helpful with the omissions items, because in going from familiar G items to omissions, the decrement in proportion of G responses was larger (.40) in group $SF_{\text{sem I}}$ than in any other group.

Table 8 here

More intensive analyses of these data indicate that this response difference among the three $\bar{G}$ categories cannot readily be explained by appealing to transfer from the native language. Grammaticality in English would be preserved despite some, but far from all, omissions, syntactical and morphological changes. Roughly speaking, the morphology proportions seem compatible with the interpretation that the subject first learns individual words. Evidently he learns only the SVO words but not the particles, because in all but one of the omissions items the missing element was a particle. This point will be covered in more detail in the section on individual item analyses. Our thesis, that the subject learns individual words as such, means that he does not merely learn stems or suffixes. For if he learned only the suffixes, he would call grammatical an utterance artifically constructed so that its suffix sequence, but not its stem sequence, occurs in the language. In our earlier notation, such an item was denoted schematically as in part III of Table 4. The low proportion of G responses to these strings (e.g., .22 on Day 1) indicates that the subject did not simply orient to the relevant elements of the sentence (suffixes and particles) and learn their syntax. Recall that three of the morphology items had a grammatical SVO stem sequence but an ungrammatical suffix sequence, whereas for the other three it was the suffix sequence that was grammatical. Resorting to our earlier notation, the error sum over $S_vV_oO_s$, $S_sV_oO_v$, and $S_oV_sO_v$ was 104, against 86 on $V_sO_vS_o$, $S_sO_vV_o$, and $O_sS_vV_o$. There is no clear difference between the types, probably because the mere presence of any pseudoword (e.g., S_v) sufficed to produce a preponderance of $\bar{G}$ responses. Such an explanation accords with our contention that words are learned as units. As expected, the morphology items to which the incorrect response G occurred most frequently were the two containing only two pseudo-words, i.e., the two with S_s.

Data from both the G and the $\bar{G}$ items show that an element of novelty in the utterance increased the probability of its being called ungrammatical. The relationship between novelty and semantics was mentioned earlier, so now we consider only the nonsemantic conditions. Referring to rows 7-9 of Table 8, the proportion of errors on $\bar{G}$ items generally decreased from familiar triples (low novelty) to new combinations of familiar words (intermediate novelty) to utterances containing a new word (high novelty). Correspondingly, with the familiar G items, rows 4-6 of the table reveal that the proportion correct was about .15 less on the high novelty items than on the low and intermediate novelty ones. Thus there was a global tendency to give the response $\bar{G}$ to unfamiliar utterances, causing the proportion correct to vary directly with novelty of $\bar{G}$ items and inversely with novelty of G items. An occurrence of a new combination is to some extent recognized as novel, and an utterance with a new word is considered even less familiar. Among $\bar{G}$ items, rote memorization seems responsible for the slight difference between the familiar utterance and new combinations of familiar words. Finally, the fact that the introduction of a new word increases the probability of a $\bar{G}$ response supports the earlier contention that orientation to individual words is an important determinant of classificatory responses.

Intonation. Here the question was how word stress, rather than overall sentence intonation, influenced judgments of grammaticality. Recall that each S, V, and O contained two syllables, and within a vocabulary set all members of the same form-class were normally stressed on the same syllable. In sets I and II, the stress patterns were 1-1-2 and 1-2-2, respectively, where the first, second, and third digits denote the stressed syllable of the subject, verb, and object, respectively. We sought to pit the syntactical and intonational features against each other and see which dominated in producing G responses. As noted in Table 4, the anomalous intonation items had a grammatical SVO word order with all words properly inflected but one or more stressed on the wrong syllable. On the other hand, the SVO intonation items were utterances with improper (e.g., VOS) word order, but the V, O, and S were stressed on the first, first, and second syllables, respectively, in set I; i.e., only the verb was correctly stressed. Note that both item types share the property that individual words are stressed on the wrong syllable; the two types differ in what property of the natural language - word order or rhythm - is preserved.

As indicated by the last two rows of Table 8, the proportion of G responses ranged from .35 to .54 on the anomalous intonation items, whereas the range on the SVO intonation items was only .10 to .18. These low proportions on the latter items imply that subjects did not simply learn a 1-1-2 rhythm and call grammatical any utterance with this rhythm. According to these data word order predominates over rhythm in producing G responses. However, even when the word order was correct, the presence of improperly stressed words was in itself sufficient to produce above-chance proportions of $\bar{G}$ responses, because Pr ($\bar{G}$) exceeded .50 for six of eight groups on Day 1 and six of seven on Day 2.

Individual item analyses. The item categories in Tables 5, 6, and 8 can be further subdivided in order to see how responses depended on the structural variables such as the paradigm (D, N, Q, NQ, P) and the particular deviation from grammaticality (e.g., what elements were omitted or transposed). Here the data are pooled over all presentation conditions and both days, but separately for each vocabulary set. Only those findings that are consistent from one set to the other need be taken seriously, because such regularity minimizes the possibility that the outcome is an artifact of the particular words in the utterance. In interpreting these data, it is helpful to refer to the item structures listed in Table 4.

Table 9 here

Table 9 summarizes the comparisons among D, N, Q, NQ, and P for the familiar G triples. Each total in the next-to-last column is based on 600 observations. Both the set I and set II subtotals support the conclusion that the rank-order from most to least difficult paradigm was N, P = Q, D = NQ. Thus one paradoxical result is that NQ, although structurally more complex than Q, (NQ is Q preceded by the negation) was correctly judged grammatical more often than was Q. Prefacing an utterance by a negation does not always increase the proportion of G responses, because an additional tally of the new paradigm items showed no more G responses to NPQ than to PQ. One hypothesis to account for the extreme difficulty of the N items might be that the D pattern is overlearned and perceptible deviations from it are judged ungrammatical. This account appears implausible, because it suggests a more general principle which implies that Q should be easier than NQ, contrary to the data. Instead, what may account for the relatively high proportion of errors on the N paradigm is the fact that only here does the object take the genitive inflection; for example in set I the O^{gen} would be heard as /a/ whereas the O^{nom} and O^{acc} appears as non-suffixed stems.

Corroborative evidence that the object (or at least the final word of the utterance) is a peculiarly salient cue comes from the anomalous intonation items. Adopting the notation that a bar above a symbol denotes stress on the wrong syllable, there were two items of each of the stress patterns $\bar{S}\bar{V}\bar{O}$, $\bar{S}\bar{V}O$, and $SV\bar{O}$. As expected, the first type yielded the most $\bar{G}$ responses (cf. Table 5). But the critical outcome is that Pr ($\bar{G}$) was .70 on $SV\bar{O}$ items, against only .40 on $\bar{S}\bar{V}O$ ones. Thus a sentence with one improper stress was judged ungrammatical much more often than one with two improper stresses, indicating that the stress on the final word of the sentence is a potent cue to grammaticality. Perhaps the stress pattern of the earlier elements is partially forgotten.

Now we turn to the question of what forms of ungrammaticality (in addition to word stress) are most readily detected as such. Of the five items with improper syntax, two had transposed particles and three had ungrammatical permutations of the elements S, V, O. One of the transpositions was an interchange of V and ЛИ in the Q paradigm while the other was an interchange of V and HE in the N paradigm. The three ungrammatical permutations of S, V, O were VOS, OSV, and SOV, presented without any particles. According to Table 9 the overall proportion of errors was .59 on transposed particles, against only .19 on SVO permutations. As a baseline for evaluating this outcome, recall that on grammatical items N and Q yielded more $\bar{G}$ responses than did D. But distortions of N and Q patterns lead to fewer $\bar{G}$ responses than distortions from the D pattern. Independent evidence that the subjects orient more to S, V, and O rather than to the particles is provided by the data from the five omissions items. The one item where the verb was omitted led to an error proportion of .39, based on all groups, sets, and days, whereas the items with omitted particles yielded .65, .66, .74, .71 errors. One omissions item uniquely provides information on the degree of morphology learning. This item is S^{nom} V O^{gen}, to which the correct response $\bar{G}$ could occur either by noting that HE has been omitted from the negative or by noting that the declarative ends in O^{acc}, not O^{gen}. The above chance error proportion of .66 demonstrates that inflections and particles were only imperfectly learned.

Hence, data from both the improper syntax items and the omissions items indicate better syntax mastery of S, V, and O than of particles or inflections. Evidently the subject first learns that the admissible orders are SVO (in D and N), VSO (in Q and NQ) and OVS (in P), then fills in the particles HE and ЛИ. If we now add the evidence from the morphology items, we can elaborate this conclusion; he integrates stems with suffixes and learns SVO, VSO and OVS as word sequences, not merely stem or suffix sequences. Thus the chain of learning events seems to be: associating S, V, O suffixes to stems, learning SVO orders and, lastly, learning particles and inflections.

Production Test.

Recall that immediately after the classification test the subject was given a series of 45 controlled sentence-production items. The subject's production was controlled in that he was to repeat the sentence he had just heard, except that one word was to be replaced by the isolated word presented immediately after the sentence. Criterion for judging a subject's production to be correct was that it be preceded and followed by the proper words, thus indicating that it had been inserted in the proper slot in the sentence. This was a lenient criterion, because neither errors earlier or later in the sentence, nor mispronunciations which did not render the words unintelligible, were penalized. Evaluation of the tapes was performed by two graduate assistants, one with four and the other with two years' training in Russian. Actually, for the criterion mentioned above, less sophisticated judges would have sufficed. As a reliability check, two untrained observers also graded several protocols, and their evaluation correlated beyond .95 with those of the two more expert judges, whose ratings in turn correlated equally high with each others. Use of trained judges enabled a more detailed classification of the type of production error, but these refinements need not concern us here.

Table 10 gives the number of errors averaged over the ten subjects per group. These quantities range from 9.3 to 17.3, corresponding to error proportions of .21 and .38. Therefore under the more effective learning conditions, subjects were able to produce sentence substitutions, erring on what seems a rather low proportion of the instances considering the fact that they had received less than two hours' training. Of course the proficiency is in part attributable to the fact that the subject only changed one word in the sentence. We would expect less facility in production tests where the subject had to, say, transform a given entire sentence or produce a sentence without a cueing sentence.[3] Kendall's rank-order correlations were computed between individual subject error frequencies on classification and on production. Most groups yielded coefficients in the range of + .20 to + .40.

Table 10 here

The production data show that proficiency is measurably higher when translations are given than when they are withheld. Evidence for this statement derives from the Day 1 comparison between groups SF_I-R_{II} and SF_{semI}, who yielded means of 16.7 and 9.3 errors, respectively (Mann-Whitney $U = 5$, two-tailed $p < .002$). Secondly, the Day 1 comparison between groups SF_I-R_{II} and SF_{II}-R_I reveals a prominent vocabulary effect ($U = 13$, two-tailed $p < .05$). The interpretation is that set I occasioned numerous pronunciation errors owing to confusions between the unvoiced stops /k/ and /t/ in the final position. Closer inspection of the data revealed that these endings, which in our set I characterized nouns and verbs, respectively, were commonly confused during production. This explanation also accounts for the fact that the influence of vocabulary set was greater on production than on the classification test. Another consequence of the vocabulary effect is that the random order conditions can be tested against only one nonrandom group; namely, the one (SF_{II}-R_I) which also had set II. This one-tailed U-test yielded $z = 1.56$, $p < .06$. Hence all three variables - semantics, vocabulary, and presentation order, influenced production more than classification. And on both tests the effect of presentation order was subordinate to those of semantics and vocabulary. Another finding confirmed by both tests was the lack of positive transfer from the Day 1 vocabulary and order to the Day 2 vocabulary and order. Improvement across days resulted only when going from vocabulary set I to set II, not when going from II to I, and hence is attributable to the set, as already noted.

Among the nonrandom orders, Table 10 reveals essentially equivalent performance under P_I-R_{II}, SC_I-R_{II}, and SF_I-R_{II}, both on Day 1 and on Day 2. Among the three corresponding groups who had R_{II} on Day 1, the 10.7 entry suggests that on Day 2 SC_I was superior to P_I and to SF_I. That this result is an artifact is seen from the fact that group R_{II}-SC_I had an unusually low error frequency even on Day 1, where they were treated identically to groups R_{II}-P_I and R_{II}-SF_I. (Of the three subjects whose data were responsible for this incongruity, post-experimental interviews disclosed that one had had extensive training in oriental languages, and one had taken several advanced linguistics and language courses. It is unlikely that other subjects' backgrounds were this complete.)

Which of the five sentence paradigms produced the most errors? Using English material and different procedures, several investigators (e.g., Miller, 1962; Mehler, 1963; Lane and Schneider, 1963) have found generally better performance in the declaratives and poorer performance with "double" transforms, such as NQ. One exception is that Lane and Schneider's subjects tended to confuse speakers of declaratives with speakers of passives. Excluding groups $SF_{II}-R_{I}$ and SF_{semI}, the error proportions pooled over all groups and sets for Q, NQ, D, P, and N were .36, .34, .29, .26, and .24 respectively. Because of the previously noted significant effect of vocabulary set, this tabulation was recalculated separately for the two sets. Both upheld the result from the pooled data; the major division was between Q and NQ as more difficult with D, P, and N as easier and no appreciable differences among these three.

The next analysis sought to discover why the Q and NQ paradigms exceeded the others in total errors. Each of the 45 sentences spoken by the experimenter was classified according to whether it required substitution of the first, second, or third word (not counting particles) and whether the word to be substituted was a subject, verb, or object. Thus each sentence was assigned to one cell of a 3 x 3 table. The following results for the set II random order conditions are typical of the overall trend. Among the object words, the D, N paradigms yielded a mean of 11.2 errors, against 12.1 for Q, NQ. (These quantities are averages over the three utterances for each paradigm and over the two paradigms). Among the verbs, the D, N paradigms were slightly superior to Q, NQ, since the error means were 14.0 and 17.7, respectively. But the major difference occurred with the subject words; 11.5

errors in D, N and 26.8 in Q, NQ. Insofar as this analysis into individual words is appropriate, the greater difficulty of Q and NQ reflects failure to make a proper substitution for the subject word.

Summary

Acquisition of the syntax of a second-language is a fruitful subject for experimentation for at least two reasons: it requires the subject to employ rule learning processes presumably akin to those investigated in concept learning studies, and it furnishes data that bear on the pedagogic merits of different instructional methods. The research reported here examined how selected task and item variables affect two measures of competency in using Russian grammar - ability to judge whether or not an utterance is grammatical and ability to produce a grammatical utterance by substituting into a given one. Task variables examined were (1) the presence vs. absence of concomitant semantic information, (2) the particular Russian words composing the three or four-word sentences, and (3) the order of presenting these grammatical sentences during training. The item variable was the syntactical and morphological structure of the test item.

The experiment consisted of three phases. First occurred acquisition, in which the subject heard a series of 160 auditorily presented grammatical sentences. He was to imitate each and learn by induction certain syntactical and morphological requirements that a grammatical utterance must satisfy. These sentences were systematically constructed from a twelve-word vocabulary, and included an equal number of declaratives, negatives, questions, negative questions, and passives. All task variables were manipulated as between-groups treatments during the acquisition phase. Four presentation orders were used, random (R), "string constituent" (SC), "substitution-in-frame" (SF), and "permutation" (P). Condition SC represented minimal paradigm variation and maximal vocabulary variation from one presentation to the next. Conversely, paradigm variation was maximized and vocabulary variation minimized in condition P. Order SF was analogous to a frequently used classroom pattern drill. A second SF group was trained and tested with a different vocabulary, and a third SF group heard the English translation after each sentence during acquisition.

The classification test for all groups involved a random series of 29 grammatical and 38 ungrammatical utterances. After hearing each, the subject said whether or not he thought it was grammatical. No feedback was given following the response. The 67 items varied from one another along several dimensions: (1) novelty - e.g., some test sentences contained a word that the subject had not heard previously; (2) the ungrammatical items were partitioned according to which one of three features - syntax, morphology, or intonation - was improper. The reason for item heterogeneity was to ascertain whether syntactical, morphological, or intonation cues dominated in determining judgments of grammaticality. On the controlled production test the subject heard a grammatical sentence followed by an isolated word. His task was to substitute the isolated word for one of the words in the sentence; the question was whether in his oral response the correct word would be replaced.

Analysis of the binary classifications showed moderate facilitation attributable to presence of translations, but no notable effect of vocabulary or presentation order. Of greater interest were the item analyses. Syntactical and intonational irregularities were usually detected by the subjects, whereas improper inflections and omissions of particles were largely overlooked. Regardless of whether an item actually was grammatical, if it contained a new word there was a pronounced increase in the proportion of judgments "ungrammatical". Regarding overall proficiency, all groups exhibited high proportions correct on items whose words and paradigms were familiar, but transfer to utterances with unfamiliar words or paradigms depressed performances appreciably.

The production test was more sensitive to all experimental treatments than was the classification test. Evaluation of the taped utterances was based on grammaticality, not on quality of production. Proportions of errors for various groups ranged from .21 to .38 after less than two hours' training. Presence of translations during acquisition significantly facilitated production. Also, one vocabulary set was significantly harder than the other, apparently because the former contained the easily confused suffixes /k/ and /t/. The superiority of condition SF to the random condition approached statistical significance, but performance differences among the nonrandom orders were slight. Clearly the effect of presentation order was overshadowed by the effects of translations and vocabulary set. It is concluded that future theoretical analyses should seek to explicate the function of semantic information, relegating presentation order factors to a secondary role for tasks such as the present one.

Footnotes

[1] These two types would be represented in set I by items the form (V + ИК , O + ИТ , S + Н) and (S + ИТ , V + Н , O + ИК), respectively.

[2] The same conclusions obtain if these items are included.

[3] And in fact pilot data, not reported here, do indicate that proportions correct average only around .30 for tests of this nature, following the same training conditions as in the text.

TABLE 1

The five sentence paradigms

PARADIGM	EXAMPLE Transcription	Translation	Structures
D	ljoščjik pomnjit kamjin	The airman remembers the fireplace.	$S^{nom} + V + O^{acc}$
N	ljoščjik nji pomnjit kamjina	The airman does not remember the fireplace.	$S^{nom} + \text{НЕ} + V + O^{gen}$
Q	pomnjit lji ljoščjik kamjin	Does the airman remember the fireplace?	$V + \text{ЛИ} + S^{nom} + O^{acc}$
NQ	nji pomnjit lji ljoščjik kamjin	Doesn't the airman remember the fireplace?	$\text{НЕ} + V + \text{ЛИ} + S^{nom} + O^{acc}$
P	kamjin pomnjica ljoščjikam	The fireplace is remembered by the airman.	$O^{nom} + V(-\text{СЯ}) + S^{inst}$
PQ*	pomnjica lji kamjin ljoščjikam	Is the fireplace remembered by the airman?	$V(-\text{СЯ}) + \text{ЛИ} + O^{nom} + S^{inst}$
NP*	kamjin nji pomnjica ljoščjikam	The fireplace isn't remembered by the airman.	$O^{nom} + \text{НЕ} + V(-\text{СЯ}) + S^{inst}$
NPQ*	nji pomnjica lji kamjin ljoščjikam	Isn't the fireplace remembered by the airman?	$\text{НЕ} + V(-\text{СЯ}) + \text{ЛИ} + O^{nom} + S^{inst}$

*Appeared as the new paradigms on the classification test.

TABLE 2

Phonemic Transcription of Suffixes

	Set I	Set II
S^{nom}	- ИК /jik/	-vowel + Р/(V)r/
S^{inst}	- ОМ /am/	- ОМ /am/
V	- ИТ */jit/	- ЛИ */jil/
O^{nom}	-vowel + Н/(V)n/	-vowel + Й /(V)j/
O^{acc}	-vowel + Н/(V)n/	-vowel + Й /(V)j/
O^{gen}	-А /a/	-А /a/

*Append - СЯ in the P, Pq, NP, and NPQ paradigms.

TABLE 3

Presentation Order for each Acquisition Condition

Permutation (P)	String Constituent (SC)	Substitution-in-Frame (SF)	Random (R)
D - $S_1V_1O_1$	D - $S_1V_1O_1$	D - $S_1V_1O_1$	All 160
P - $O_1V_1S_1$	D - $S_2V_2O_2$	D - $S_1V_1O_2$	in
N - $S_1V_1O_1$	D - $S_3V_3O_3$	D - $S_1V_2O_3$	random
Q - $V_1S_1O_1$	all 32 D,	D - $S_1V_2O_4$	order.
NQ - $V_1S_1O_1$	then 32 P,	All 32 D, then	
Then recycle with a different SVO triple	then 32 N, etc.	32 P, then 32 N, etc.	

TABLE 4

Classification Test Items*

I. Varying novelty:

	No. of grammatical (G)	Description	No. of ungrammatical ($\bar{G}$)	Description
Familiar SVO triple	10	2 ea. of D,N,Q,NQ,P	6	Improper syntax. Each set had one ea. of VOS, OSV, and SOV:, one ea. of Q(V, ЛИ), N(V,HE), and P(O + СЯ).
Familiar words	5	1 ea. of D,N,Q,NQ,P	6	
One new word	5	1 ea. of D,N,Q,NQ,P	6	
New paradigm	9	3 ea. of NP,PQ,NPQ	-	
	29		18	

II. $\bar{G}$ intonation items (familiar SVO triples):

Anomalous stress	6	Proper syntax and morphology, one or more of SVO stressed on wrong syllable; 2 each of $\bar{S}\bar{V}\bar{O}$, $\bar{S}\bar{V}O$, $SV\bar{O}$.
SVO stress	3	Improper syntax (VOS, OSV, SOV) but SVO stress pattern.
	9	

III. Other $\bar{G}$ items (familiar SVO triples):

Omissions	5	One with each of D(-V), N(-HE), Q(- ЛИ), NQ(-HE, ЛИ), and P(- СЯ).
Morphology	6	One each of $S_vV_oO_s$, $S_oV_sO_v$, $S_sV_oO_v$, $V_sO_vS_o$, $O_sS_vV_o$, and $S_sO_vV_o$.
	11	

---

*See text for explanation of notation.

TABLE 5

Day 1 Error Frequencies on Grammatical (G) and Ungrammatical ($\bar{G}$) Items*

Item Type	No. of Items (N)	SF'_{semI}	SF_{I}	SF_{II}	R_{IImean}
G, familiar paradigm	20	38	22	46	36.0
G, new paradigm	9	57	48	40	52.3
$\bar{G}$, all types	32	72	112	109	109.0
anomalous intonation**	6	28	38	48	44.3
	67				

---

* To obtain corresponding proportions, divide each entry by 10 N.

** Entries are frequencies of judgment "ungrammatical".

TABLE 6

Error Frequencies for Nonrandom Conditions with Vocabulary Set I*

Item Type	Day 1			Day 2**		
	P_I	SC_I	SF_I	P_I	SC_I	SF_I
G, familiar paradigm	21	20	21	31	29	27
G, new paradigm	55	76	48	51	68	53
$\bar{G}$, all types	118	116	112	116	76	97
anomalous intonation	28	36	38	33	41	37

---

*The footnotes to Table 5 apply here as well.

**These three groups had condition R_{II} on Day 1.

TABLE 7

Error Frequence per Subject on Classification Test (61 items)

Group		Day 1		Day 2	
Day 1	Day 2	mean	range	mean	range
R_{II} -	P_I	19.0	11-24	19.8	12-25
R_{II} -	SC_I	19.7	16-22	17.3	12-21
R_{II} -	SF_I	20.2	13-27	17.7	9-25
SF_{II} -	R_I	19.5	13-25	17.7	16-21
P_I -	R_{II}	19.4	15-23	20.1	15-27
SC_I -	R_{II}	21.2	18-25	20.9	16-26
SF_I -	R_{II}	18.2	14-23	20.4	14-26
SF_{semI}		16.7	12-26	-	-

TABLE 8

Proportion of Judgments "grammatical" as function of Item Category

	Day 1			Day 2		
	Gps.1-6*	SF_{semI}	$SF_{II}-R_I$	Gps.1-6*	SF_{semI}	$SF_{II}-R_I$
G items (familiar vocabulary)						
Omission	.65	.50	.62	.63	-	.62
Improper syntax	.38	.23	.35	.39	-	.38
Improper morphology	.23	.23	.17	.22	-	.08
G items, effect of novelty						
Familiar utterance	.91	.90	.81	.88	-	.91
New combination	.90	.90	.82	.93	-	.84
New word	.72	.64	.64	.73	-	.72
$\bar{G}$ items, effect of novelty						
Familiar utterance	.38	.23	.35	.39	-	.38
New combination	.35	.18	.37	.36	-	.25
New word	.30	.20	.32	.32	-	.30
Correct syntax, anomalous intonation	.35	.53	.47	.35	-	.54
Improper syntax, SVO intonation	.11	.13	.10	.18	-	.10

*Pooled data from all groups other than SF_{semI} and $SF_{II}-R_I$.

TABLE 9

Error Frequencies on Classification Test Structures

	Set I			Set II				
G, old paradigm	old sentences plus new combinations	new word	$\sum$	old sentences plus new combinations	new word	$\sum$	Total	p
D	21	16	37	22	11	33	70	.12
N	26	30	56	39	34	73	129	.22
Q	22	22	44	27	23	50	94	.16
NQ	22	7	29	17	22	39	68	.11
P	26	27	53	23	22	45	98	.16
Anomalous Intonation								
$\bar{S}\bar{V}\bar{O}$	--	--	130	--	--	124	254	.85
$\bar{S}\bar{V}O$	--	--	46	--	--	73	119	.40
$SV\bar{O}$	--	--	106	--	--	104	210	.70
Improper* Syntax								
Transposed particle	95.5	35.5	131.0	102.0	30.5	132.5	263.5	.59
VOS, OSV, SOV	26.0	9.7	35.7	42.0	9.7	51.7	87.5	.19

*Normalized to adjust for presentation frequency differences between the two rows.

TABLE 10

Substitution Production Test: Mean Error Frequency on 45 Items

Group	Day 1	Day 2
$R_{II} - P_{I}$	15.2	17.3
$R_{II} - SC_{I}$	10.1	10.7
$R_{II} - SF_{I}$	14.2	15.2
$SF_{II} - R_{I}$	10.0	16.3
$P_{I} - R_{II}$	15.3	11.3
$SC_{I} - R_{II}$	16.3	10.5
$SF_{I} - R_{II}$	16.7	10.5
SF_{semI}	9.3	-

References

Atkinson, R. C. and Crothers, E. J. A comparison of paired-associate learning models having different acquisition and retention axioms. J. math. Psychol., 1964, 1, 285-315.

Calfee, R. C. and Atkinson, R. C. Paired-associate models and the effect of list length. J. math. Psychol., 1965, 2, 255-267.

Chomsky, N. Review of Verbal Behavior, B. F. Skinner, Language, 1959, 35, 26-58.

Crothers, E. J. Optimal presentation orders for items from different categories. Tech. Rep. No. 71, Institute for Mathematical Studies in the Social Sciences, Stanford Univer., 1965.

Lane, H. and Schneider, B. Some discriminative properties of syntactic structures. J. verb. Learn. verb. Behav., 1963, 2, 457-461.

Mehler, J. Some effects of grammatical transformations on the recall of English sentences. J. verb. Learn. verb. Behav., 1963, 2, 346-351.

Miller, G. Some psychological studies of grammar. Amer. Psychol., Nov. 1962, 748-762.

Norman, M. F. A two-phase model and an application to verbal discrimination learning. In R. C. Atkinson

(Ed.), Studies in mathematical psychology. Stanford: Stanford Univer. Press, 1964, 173-187.

Siegel, S. Nonparametric statistics. New York: McGraw-Hill, 1956.

Suppes, P. Problems of optimization in learning a list of simple items. Tech. Rep. No. 57, Institute for Mathematical Studies in the Social Sciences, Stanford Univer., 1963.

Underwood, B. J. and Schulz, R. W. Meaningfulness and verbal learning. Chicago: Lippincott, 1960.

APPENDIX A

Grammar Experiment Instructions

At start of experiment.

We are doing this experiment to see how well people can learn the correct grammatical forms of Russian sentences. You will hear Russian sentences, one at a time, through your earphones. Each time you hear a sentence, please repeat it aloud as well as you can. Speak into the microphone so that we can record your imitation of the native speaker. Please say the sentence before the native speaker goes on to the next sentence. Naturally your pronounciation will be poor at first, but you will gradually improve. All your training will be by imitating sentences that you hear - no English translations of these sentences will be given.

After you have imitated a number of sentences, we will move on to the second part of the experiment. There you will hear Russian utterances, one at a time. Your task will be to indicate after each utterance whether you think it is a grammatically correct sentence or a grammatically incorrect sentence. We will give you more specific instructions when this second part of the experiment begins. In the final part of the experiment, you will be asked to produce Russian sentences instead of merely imitating and judging them.

Classification Test.

Now, when you hear an utterance, do not imitate it aloud. Instead, on the test sheet you have been given write opposite the appropriate number a plus if you think the utterance is grammatically correct or zero if you think it is not grammatically correct. Several examples from English will clarify what we mean by an utterance that is not grammatically correct. Take the correct sentence, "The man rides a bus". One type of incorrect sentence has an improper word order; for example, "Man the rides a bus". Another type of incorrect sentence has the right word order but the wrong forms. For example: "The man ride a bus". A second incorrect example of this type is, "The man rides a busses". Another type of ungrammatical sentence has a word omitted, for example, "The man rides bus". After you hear a sentence, if you think the word order, or word endings are incorrect or a word is omitted, put a zero by the sentence number. But if you think the sentence is grammatically correct, then put a plus by the sentence number. Please keep up with the numbers. Guess if you are not sure. Numbers 22 and 44 of your test sheet will serve as check points. The tape will be stopped at these two numbers and you will be asked whether you have marked your sheets thus far.

Production Test Part B

In this exercise you will be given a Russian sentence. Then you will be given one word. You are asked to make a correct new sentence by replacing a word of the sentence you heard with the word you were given. It is up to you to decide which word of the sentence you should replace. Again we illustrate with an example from English. If the sentence "Where is the book" were followed by "toy", the new correct sentence would be "Where is the toy?"

APPENDIX B

Russian Words Used in Grammar Experiment

Subjects (S)	Verbs (V)	Objects (O)
1. МА́ЛЬЧИК boy	1. ПО́МНИТ remembers	1. КУВШИ́Н pitcher
2. ЛЁТЧИК flyer	2. ЧИ́СТИТ cleans	2. ДИВА́Н sofa
3. ПЛО́ТНИК carpenter	3. МЕ́РИТ measures	3. КАМИ́Н fireplace
4. МЕ́ЛЬНИК miller	4. ВИ́ДИТ sees	4. ЭКРА́Н screen

SET II

1. А́ВТОР author	1. ЧИНИ́Л repaired	1. ТРАМВА́Й streetcar
2. ПО́ВАР cook	2. ЦЕНИ́Л valued	2. МУЗЕ́Й museum
3. КУ́ЧЕР driver	3. ЛЮБИ́Л liked	3. САРА́Й barn
4. ФЕ́ЛЬДШЕР assistant	4. ДАВИ́Л crushed	4. ШУГА́Й blouse

particles: НЕ ("not") and ЛИ (denotes interrogative).

APPENDIX B

continued

New Words Introduced on Classification Test

SET I

Subjects (S)	Verbs (V)	Objects (O)
1. БРА́ТИК brother	1. КРА́СИТ paints	1. КАФТА́Н tunic
2. СКО́ТНИК cowboy	2. ПО́РТИТ spoils	2. СТАКА́Н glass

SET II

Subjects (S)	Verbs (V)	Objects (O)
1. ДИ́КТОР announcer	1. ГУБИ́Л ruined	1. ЛИЦЕ́Й school
2. МА́КЛЕР broker	2. ХВАЛИ́Л praised	2. УЛЕ́Й beehive